

A Concise Forwarding Information Base for Scalable and Fast Flat Name Switching

Ye Yu

University of Kentucky
ye.yu@uky.edu

Djamal Belazzougui

CERIST
dbelazzougui@cerist.dz

Chen Qian

UC Santa Cruz
cqian12@ucsc.edu

Qin Zhang

Indiana University Bloomington
qzhangcs@indiana.edu

Abstract

Forwarding information base (FIB) scalability is a fundamental problem of numerous new network architectures that propose to use location-independent network names. We propose Concise, a FIB design that uses very little memory to support fast query of a large number of location-independent names. Concise makes use of minimal perfect hashing and relies on the SDN framework and supports fast name classification. Our conceptual contribution of Concise is to optimize the memory efficiency and query speed in the data plane and move the relatively complex construction and update components to the resource-rich control plane. We implemented Concise on three platforms. Experimental results show that Concise uses significantly smaller memory to achieve faster query speed compared to existing FIBs for flat name switching.

1. INTRODUCTION

Significant efforts have been devoted to the investigation and deployment of new network architectures in order to simplify network management and to accommodate emerging network applications. Though different proposals of new network architectures focus on a wide range of issues, one consensus of most new network designs is the separation of network identifiers and locators [37], which are combined in IP addresses. Instead of IP, flat-name or namespace-neutral architectures have been proposed to provide persistent network identifiers. A flat or location-independent namespace has no inherent structure and hence imposes no restrictions to referenced elements [4].

The Salter's taxonomy of network elements [37] is one of the early proposals that suggest the separation of network identifiers and locators. We summarize an (incomplete) list of reasons of using flat or location-independent names in proposed network architectures:

- To simplify network management, pure layer-two Ethernet is suggested to interconnect large-scale enterprise and data center networks [14, 19, 24, 35, 38, 40], where MAC addresses are identifiers.
- Flat network identifiers have been suggested by various works to support host mobility and multi-homing, includ-

ing HIP [31], Layered Naming Architecture [4], DONA [26] and MobilityFirst [36].

- AIP [2] and XIA [33] apply flexible addressing to ensure trustworthy communication.
- The core network of Long-Term Evolution (LTE) needs to forward downstream traffic according to the Tunnel End Point Identifier (TEID) of the flows [46].
- Software Defined Networking (SDN) [29] uses matching of multiple fields in packet header space to perform fine-grained per-flow control. Flow IDs can also be considered names, though they are not fully flat.

The most critical problem caused by location-independent names is *Forwarding Information Base (FIB) explosion*. A FIB is a data structure, typically a table, that is used to determine the proper forwarding actions for packets, at the data plane of a network forwarding device (e.g. switch or router). Forwarding actions include sending a packet to a particular outgoing interface and dropping the packet. Determining proper forwarding actions of the names in a FIB is called name switching. Unlike IP addresses, location-independent names are difficult to aggregate due to the lack of hierarchy and semantics. The increasing population of network hosts results in huge FIBs and their continuing fast growth.

On the other hand, the increasing line speed requires the capability of fast forwarding. To support multiple 10G Ethernet links, a FIB may need to perform hundreds of millions of lookups per second. Existing high-end switch fabrics use fast memory, such as TCAM or SRAM, to support intensive FIB query requests. However, as discussed in many studies [11, 22, 35, 43, 44], fast memory is expensive, power-hungry, and hence very limited on forwarding devices. Therefore, *achieving fast queries with memory-efficient FIBs is crucial for the new network architectures that rely on location-independent names*. If FIBs are small and increase very little with network size, network operators can use relatively inexpensive switches to build large networks and do not need frequent switch upgrade when the network grows. Hence, the cost of network construction and maintenance can be significantly reduced. For software switches, small FIBs are also important to fit into fast memory such as cache.

In this paper, we present a new FIB design called Concise. It has the following properties.

1. Compared to existing FIB designs for name switching, Concise supports *much faster name lookup* using *significantly smaller memory*, shown by both theoretical analysis and empirical studies.
2. Concise can be efficiently updated to reflect network dynamics. A single CPU core is able to perform millions of network updates per second. Concise makes the control plane highly scalable.
3. Concise guarantees to return the correct forwarding actions for valid names. It is *not* probabilistic like those using Bloom filters [30, 43].

Concise is based on a data structure called Othello, which uses existing theoretical studies on minimal perfect hashing [5, 8, 12, 28]. In this paper, minimal perfect hashing is applied to the FIB design by accommodating the SDN framework and network dynamics. Othello represents a new data structure model called Polymorphic Data Structure (PDS). A PDS performs different functionalities on heterogeneous platforms. As a data structure, Othello supports both query and update (addition/deletion of names). In the resource-limited switches (data plane), Othello only includes the query component and is optimized for memory efficiency and query speed. The construction and update components are moved to the resource-rich control plane. Construction and updates of Othello are computed in the control plane and transmitted to the data plane part via a standard API such as OpenFlow [29]. Concise is designed for name switching, so it does *not* support IP prefix matching.

Some recently proposed forwarding engines, such as CuckooSwitch [47] and ScaleBricks [46], are specifically designed for customized and high-end hardware platforms. Unlike them, Concise is a **portable solution** with no restriction on the underlying computing platform. Concise can be used in either software or hardware switches. We target on a generalized solution and apply no platform-specific optimization. We have implemented Concise in three different computing environments: memory mode, CLICK Modular Router [25], and Intel Data Plane Development Kit (DPDK) [17]. The experiments conducted on an ordinary commodity desktop computer show that Concise uses only few MBs of memory to support hundreds of millions name queries per second, when there are millions of names.

The rest of this paper is organized as follows. Sec. 2 presents related work. We introduce the overview of Concise in Sec. 3. We present the Othello data structure in Sec. 4 and the system design in Sec. 5. We present the system implementation and experimental results in Sec. 6. Sec. 7.1 discusses a few related issues. Finally, we conclude this work in Sec. 8.

2. RELATED WORK

Location-independent network names. Separating network location from identity has been proposed and kept repeating for over two decades. Numerous network architectures appear in the literature that suggest this concept. We discussed in Sec. 1 a number of new network architecture [2, 4, 14, 19, 24, 29, 31, 33, 35, 36, 38, 40] and their reasons of using location-independent or flat names, ranging from incorporating mobility and multihoming to trustworthy communication. A location-independent name can be a MAC address, a tuple consisting of several packet header fields [23], a file name [18, 45], a TEID [46], etc. To route packets for flat names, ROFL [10] and Disco [39] propose to use compact routing to achieve scalability and low routing stretch. ROME [35] is a routing protocol for layer-two networks that uses greedy routing whose routing table size is independent of network size. PIE [16] proposes to use isometrics embedding for scalable routing. Concise is a forwarding structure and does not deal with routing.

FIB scalability. Hashing is a typical approach to reduce the memory cost of FIBs for name-based switching. The use of Bloom filters [7] has been proposed for some scalable FIB designs [30, 43]. However, they may forward packets incorrectly due to the false positives in Bloom filters, causing forwarding loops and bandwidth waste. Wang *et al.* [41] uses GPU to accelerate name lookup in Named Data Networks. For IP lookups, SAIL [42] and Portire [3] demonstrate desirable throughput for IPv4 FIB queries. However, their performance will be challenged with flat names, which result in larger FIBs. CuckooSwitch [47] and ScaleBricks [46] use carefully revised Cuckoo hash tables [34] to reach desirable performance on specific high-end hardware platforms. ScaleBricks also makes use of a memory-efficient data structure SetSep [13] to partition a FIB to different nodes in a cluster, it does not store the names as well. We provide a comprehensive comparison of Cuckoo hashing, SetSep, and Concise in Sec. 7.2.

Minimal perfect hashing. The data structure proposed in this work, Othello, is inspired by the studies on minimal perfect hashing that minimizes the output space of perfect hash functions. In particular, Majewski *et al.* [28] invent the MWHC algorithm, which is able to generate order-preserving minimal perfect hash functions using a r -graph. MWHC is also presented as Bloomier Filters in [6]. The differences between Othello and these two studies [6, 28] include: (1) Both MWHC and Bloomier Filters are designed for static scenarios and they do not support frequent updates like Othello does. (2) Othello has the unique contribution in introducing the concept of polymorphic data structure (PDS) by accommodating the SDN framework. (3) Othello is a practical and concrete design that works for real network conditions. Othello aims to support fast flat name switching, while MWHC is for finding minimum perfect hash functions [28] and Bloomier Filter is designed for approximate evaluation queries [6]. A recent study [32] proposes a hash table with small memory size. However, the presen-

ted query throughput is orders of magnitude smaller than Concise. Meanwhile, the evaluation of its update scheme is not presented [32].

3. DESIGN OVERVIEW

3.1 Network Model

Consider a network of n end hosts, each of which is identified by a unique name. The hosts are interconnected by SDN-enabled switches. A logically central controller is responsible of network management tasks such as deciding routing paths of packets. Each switch includes a FIB. The controller communicates with each switch to install and update the switch FIB.

Each packet header includes the name of the destination host. Upon receiving a packet, the switch decides the outgoing port to which to forward the packet based on the destination name. For a packet with destination name k , the FIB returns an integer i representing a forwarding action A_i . An action A_i could be forwarding the packet to a specified port or dropping the packet.

We assume the controller knows the set S of all names in the network. In addition, Concise only accepts queries of valid names, i.e., $k \in S$. We assume that firewalls or similar network functions are installed at ingress switches to filter packets whose destination names do not exist. More discussion about eliminating invalid names is presented in Sec. 7.3.

A typical example of this model is a large-scale data center network using Ethernet [14, 24, 35, 40], where names are MAC addresses.

3.2 Data Structure Model

Concise makes use of a data structure *Othello*, one of our main contributions. An *Othello* classifies names to two disjoint sets. It can be extended to an advanced structure that classifies names to d disjoint sets, each of which represents a forwarding action.

We introduce a new model of data structures called Polymorphic Data Structure (PDS). A PDS performs different functionalities on heterogeneous platforms.

Othello exists in both the switches (data plane) and the controller (control plane). It has two different structures in the data plane and control plane:

- ***Othello* query structure** implemented in a switch is the FIB. It only performs name queries. The memory efficiency and query speed is optimized and the update component is removed.
- ***Othello* control structure** implemented in the controller maintains the FIB as well as other information used for FIB construction and updates, such as the routing information base (RIB). Its memory and computational cost is higher than that of the query structure, but still very efficient on a resource-rich controller.

Upon network dynamics, the control structure computes the updated FIBs of the affected switches. The modification is

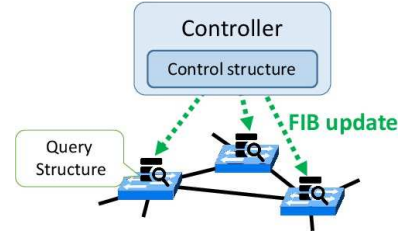


Figure 1: Network Overview

sent from the controller to each switch via a standard API such as OpenFlow [29].

The space complexity of *Othello* is αn bits, where α is a small constant and smaller than 5.72 in our implementation. Each query requires computing two hash values and two memory accesses. The extension of *Othello* to support d forwarding actions uses $\alpha \lceil \log_2 d \rceil n$ bits. Each query still only requires computing two hash values and two memory accesses.

4. Othello DATA STRUCTURE

In this section, we describe the *Othello* data structure. The static version of *Othello* is inspired by the design of MWHC minimal perfect hashing [28] and MWHC is also used in the Bloomier filter [6]. The basic function of a FIB is to classify all names into multiple sets, each of which represents a forwarding action. Let S be the set of all names. An *Othello* classifies names into two disjoint sets X and Y : $X \cup Y = S$ and $X \cap Y = \emptyset$. *Othello* can be extended to classify names into d ($d > 2$) disjoint sets, serving as a FIB with d actions.

4.1 Othello query structure

The *Othello* query structure is implemented in each switch. It supports operation $\text{query}(k)$. For a name k , it computes $\tau(k) \in \{0, 1\}$. If $k \in X$, $\tau(k) = 0$. If $k \in Y$, $\tau(k) = 1$. If $k \notin S$, $\text{query}(k)$ returns 0 or 1 arbitrarily.

We present the formal definition of the *Othello* query structure as follows. The *Othello* query structure is a six-tuple $\langle m_a, m_b, h_a, h_b, a, b \rangle$.

- Integers m_a and m_b , describing the size of *Othello*.
- A pair of uniform random hash functions $\langle h_a, h_b \rangle$, mapping names to $\{0, 1, \dots, m_a - 1\}$ and $\{0, 1, \dots, m_b - 1\}$, respectively.
- Bitmaps a and b . The lengths are m_a and m_b respectively.

For a name k , the query result $\tau(k)$ is computed by:

$$\tau(k) = a[h_a(k)] \oplus b[h_b(k)]$$

Here, \oplus is the *exclusive or* (XOR) operation. In other words, if $k \in X$, $a[h_a(k)] = b[h_b(k)]$; if $k \in Y$, $a[h_a(k)] \neq b[h_b(k)]$.

Storing this six-tuple takes $2m + O(1)$ bits memory space.¹ The time cost for each query of *Othello* is equal to the sum

¹ m_a, m_b, h_a , and h_b are stored as fixed-width integers.

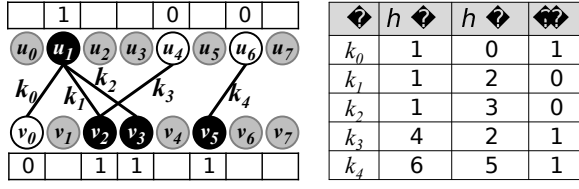


Figure 2: Example of 0thello. Left: Bipartite graph G and bitmaps a and b . Right: five names $k_0, k_3, k_4 \in X$ and $k_1, k_2 \in Y$; the hash values and $\tau(k)$ values.

of the cost of computing two hash values, twice memory accesses for the two bitmaps, and one XOR arithmetic operation.

4.2 Othello Control Structure

The 0thello control structure is maintained by the controller. It supports the following operations.

- **construct(X, Y):** Construct an 0thello for two name sets X and Y .
- **addX(k) and addY(k):** add a new name k into the set X or Y .
- **alter(k):** For a name $k \in X \cup Y$, move k from set X to Y or from Y to X . After this operation, the query result $\tau(k)$ is changed.
- **delete(k):** For a name $k \in X \cup Y$, remove k from set X or Y .

The 0thello control structure O is a seven-tuple $O = \langle m_a, m_b, h_a, h_b, a, b, G \rangle$. The first six elements have same definitions as those in the query structure.

G is a bipartite graph $G = (U, V, E)$. G is used to decide the values in a and b , so that the query $\tau(k)$ returns the correct result. In G , $|U| = m_a$, $|V| = m_b$. A vertex $u_i \in U$ ($0 \leq i < m_a$) or $v_j \in V$ ($0 \leq j < m_b$) corresponds to bit $a[i]$ or $b[j]$. Each edge in E represents a name. There is an edge $(u_i, v_j) \in E$ if and only if there is a name $k \in S$ such that $h_a(k) = i$ and $h_b(k) = j$. Each vertex that is associated with at least one edge is colored as either *white* or *black*. For a white vertex, the corresponding bit is set to 0. For a black vertex, the bit is 1. For an uncolored vertex, we do not care about the value of the corresponding bit because it does not affect any $\tau(k)$ value for $k \in S$.

Given two disjoint name sets X and Y and two hash functions h_a and h_b , the two bitmaps can be easily determined if G is *acyclic*. (See Sec. 4.3.1.) Fig. 2 shows an 0thello example of 5 names and $m_a = m_b = 8$. The left figure shows the graph G and the bitmaps a and b . The right table shows all names and their hash values and $\tau(k)$ indicating whether they belong to X or Y . Every edge $(u_{h_a(k)}, v_{h_b(k)})$ for name k is marked in G . The uncolored vertices are marked gray.

Note that, the edges of G are only determined by $S = X \cup Y$

and the hash function pair $\langle h_a, h_b \rangle$. If we find G to be cyclic for a given S and a pair $\langle h_a, h_b \rangle$, we shall use another pair $\langle h_a, h_b \rangle$ to make G acyclic. We show that for a randomly selected pair of hash functions $\langle h_a, h_b \rangle$, the probability of G to be acyclic is very high:

THEOREM 1. *Given set of names $S = X \cup Y$, $n = |S|$. Suppose h_a, h_b are randomly selected from a family of random hash functions. $h_a : S \rightarrow \{0, 1, \dots, m_a\}$, $h_b : S \rightarrow \{0, 1, \dots, m_b\}$. Then the generated bipartite graph G is acyclic with probability $\sqrt{1 - c^2}$ when $n \rightarrow \infty$, where $c = \frac{n}{\sqrt{m_a m_b}}$, $c < 1$.*

Theorem 1 can be proved from the existing results of minimal perfect hashing [9, 12]. When G is acyclic, we say $\langle h_a, h_b \rangle$ is a *valid hash function pair* for S .

The space complexity of 0thello control structure is $O(m_a + m_b)$, a very small constant times n . Here, m_a and m_b is determined during 0thello construction. We recommend their values as follows: m_a and m_b be powers of 2, meanwhile, m_a be the smallest value such that $m_a \geq 1.33n$, and m_b be the smallest value such that $m_b \geq n$. As a result, we have $2.67n \leq m_a + m_b < 4n$ and $0.5 < c < 0.75$.

Next, we show that the expected time to construct an 0thello with n names is $O(n)$ and the expected time to add/delete/alter a name is $O(1)$.

4.3 Othello Algorithms

4.3.1 Construction

The **construct** operation takes the input of two sets of names X and Y . The output is a 0thello $O = \langle m_a, m_b, h_a, h_b, a, b, G \rangle$. It consists of two phases.

(1) Deciding hash function pair

In this phase, 0thello finds a valid hash function pair $\langle h_a, h_b \rangle$ and computes the bipartite graph G . We assume there are many candidate hash functions and will discuss the implementation in Sec. 5.2. At each round, two hash functions are chosen randomly and G is accordingly generated. We use Depth-First-Search (DFS) on G to test whether it includes a cycle, which takes $O(n)$ time. The order in which the edges are visited during the DFS, i.e., the DFS order of the edges, is recorded to prepare for the second phase. Note that, if two or more names generate edges with the same two endpoints, we consider there is a cycle. If G is cyclic, the algorithm will select another pair of hash functions until an acyclic G is found.

(2) Computing bitmaps.

In this phase, 0thello assigns values for the two bitmaps a and b . First, the values in a and b are marked as undefined. Then, we execute the followings for each $e = (u_i, v_j)$ in the DFS order of edges: Let k be the name that generates e . If none of $a[i]$ and $b[j]$ has been assigned, let $a[i] \leftarrow 0$ and $b[j] \leftarrow \tau(k)$. If there is only one of $a[i]$ and $b[j]$ has been assigned, we can always assign an appropriate value to the other one, such that $a[i] \oplus b[j] = \tau(k)$. As G is acyclic, following the DFS order, we will never see an edge such that both $a[i]$ and $b[j]$ have values.

Complexity Analysis. For the first phase, the expected number of rounds to find an acyclic G is $\frac{1}{\sqrt{1-c^2}} \leq 1.51$ when $c < 0.75$. The time complexity is $O(n)$ in each round. The second phase takes $O(n)$ time to visit n edges and assign values of a and b . Hence, the total expected time of construct is $O(n)$.

4.3.2 Name addition

To add a name k to X or Y , the graph G and two bitmaps should be changed in order to maintain the correct result $\tau(k)$. An acyclic graph G can be decomposed into connected components. A connected component cc of G is a subgraph of G : any two vertices in cc are connected by an path of edges, and any vertex in cc is connected to no additional vertices in the supergraph. A connected component may contain only one isolated vertex, i.e., a vertex with no adjacent edge.

The algorithm first computes the edge $e = (u, v)$ to be added to G for k , $u = u_{h_a(x)}$, $v = v_{h_b(x)}$. e must fall in one of the following cases.

Case I: u and v belong to the same connected component. Adding e to G will introduce a cycle. In this case, we have to re-select a hash function pair $\langle h_a, h_b \rangle$ until a valid hash function pair is found for the new name set $S \cup \{k\}$. The construct algorithm is used to perform this process.

Case II: u and v are in two different connected components. Combining the two connected components and the new edge, we have a single connected component that is still acyclic. As discussed in Sec. 4.3.1, it is simple to find a valid coloring plan for an acyclic connected component. Hence, the values of a and b can also be set properly. In fact, at least one of the two connected components can keep the existing colors.

Complexity Analysis.

The *susceptibility* $E[|cc|]$ of a graph is defined as the expected size of the connected component containing a randomly chosen vertex. For general random graphs with $2m$ nodes and $n = pm$ edges, $E[|cc|]$ only depends on p , $E[|cc|] \rightarrow \frac{1}{1-p}$ when $n \rightarrow \infty$ [21]. Empirical results show that Othello bipartite graphs have a similar property: For bipartite graphs, $E[|cc|] < \frac{1}{1-c} + 0.0016$ with confidence level 99%, where $c = \frac{n}{\sqrt{m_a m_b}}$. Hence, we use $\frac{1}{1-c}$ to estimate $E[|cc|]$. Note that $c < 0.75$, $E[|cc|] \leq 4$, which is a small constant.

We also present the following proposition. Its proof and validation are omitted due to space constraints.

PROPOSITION 2. *For a name addition, the probability of falling in Case I is at most $\frac{3}{8n}E[|cc|]$.*

In Case I of add, the construct algorithm is executed in $O(n)$ time. In Case II, the values correspond to vertices in one component is updated in $O(E[|cc|])$ time. Hence, the expected time complexity is $\frac{3}{8n}E[|cc|] \cdot O(n) + (1 - \frac{3}{8n}E[|cc|]) \cdot O(E[|cc|]) = O(1)$.

Othello size growth. After adding a name into Othello, $n = |S|$ grows and may violate $m_a \geq 1.33n$ and $m_b \geq n$. How-

ever, Othello works correctly as long as G is acyclic, even $m_a < 1.33n$ or $m_b < n$. Hence, Othello do not deal with the requirement on m_a and m_b explicitly for additions. Although the $E[|cc|]$ value may grow as more names are added to Othello, it is always smaller than 15 in our experiments. In conclusion, we confirm that the expected time to add a name to Othello is $O(1)$.

When adding a new name falling in Case I, the construct algorithm is executed. In such case, the values of m_a and m_b will be updated, which guarantees $m_a \geq 1.33n$ and $m_b \geq n$.

4.3.3 Set change for a name

The goal of $\text{alter}(k)$ for a name $k \in X \cup Y$ is to move k from X to Y (or from Y to X). The bitmaps a and b should be modified so that $\tau(k)$ is changed from 0 to 1 (or from 1 to 0). Note that the graph G does not change during $\text{alter}(k)$. It is only necessary to change the coloring plan of the connected component including the edge $e = (u_{h_a(k)}, v_{h_b(k)})$. One approach is to “flip” the colors of all vertices at one side of e , i.e., to change 0 to 1, and to change 1 to 0. The expected time cost is $O(E[|cc|]) = O(1)$.

4.3.4 Name deletion

$\text{delete}(k)$ can be done by simply removing the edge $(u_{h_a(k)}, v_{h_b(k)})$ in the graph G . The bitmaps a and b are not modified because the values of $\tau(k)$ after deleting k do not matter any more. The time complexity is $O(1)$.

4.4 Summary of Othello Properties

An Othello is decomposed to a query structure running in the data plane and a control structure in the control plane. The query structure uses no more than $4n$ bits for n names. Every query takes a small constant time including computing two hashes and two memory accesses. The control structure uses $O(n)$ bits. The expect time complexity is $O(n)$ for construction and $O(1)$ for name addition, deletion, and set change. Note that *the distribution of names in X and Y has no impact to the space and time cost of Othello*, because G only depends on S and $\langle h_a, h_b \rangle$.

In Sec. 5.1, we demonstrate the extension of Othello. It classifies names to $d > 2$ disjoint sets, while still requires small memory and constant query time.

5. SYSTEM DESIGN OF Concise

This section presents the design of Concise based on the Othello data structure.

5.1 Extension of Othello for Name Switching

The extension of Othello to support classification for more than two sets is called a Parallel Othello Group (POG). An l -POG is able to classify names into 2^l disjoint sets. It serves as a FIB with 2^l forwarding actions.

Let $Z_0, Z_1, \dots, Z_{2^l-1}$ be the 2^l disjoint sets of names for an l -POG. Let $S = Z_0 \cup Z_1 \cup \dots \cup Z_{2^l-1}$. For any name $k \in S$, $\text{query}(k)$ of the l -POG returns an l -bit integer $\tau(k) \in$

$\{0, 1, \dots, 2^l - 1\}$. $\tau(k)$ is the index of the set which contains k , i.e., $k \in Z_{\tau(k)}$.

The idea of POG is as follows. Consider l 0thello control structures O_1, O_2, \dots, O_l . Each O_i classifies keys in set X_i and Y_i ($1 \leq i \leq l$), where X_i and Y_i satisfies:

$$X_i = \bigcup_{(j \bmod 2^i) < 2^{i-1}} Z_j; \quad Y_i = \bigcup_{(j \bmod 2^i) \geq 2^{i-1}} Z_j.$$

Let $\tau_i(k)$ be the query result of O_i for name k . Consider the l -bit integer $((\tau_l(k)\tau_{l-1}(k) \cdots \tau_1(k))_2)$. The i -th least significant bit $\tau_i(k) = 0$ if and only if $k \in X_i$. Meanwhile, $Z_{\tau(k)} \subset X_i$ if and only if $(\tau(k) \bmod 2^i) < 2^{i-1}$, this indicates that the i -th least significant bit of $\tau(k)$ is 0. Hence, the i -th least significant bit of $\tau(k)$ equals to $\tau_i(k)$. i.e.,

$$\tau(k) = ((\tau_l(k)\tau_{l-1}(k) \cdots \tau_1(k))_2)$$

For each i ($1 \leq i \leq l$), $X_i \cup Y_i = S$. i.e., the l 0thellos share the same S . Recall that the edges in G is determined by only $S = X \cup Y$ and $\langle h_a, h_b \rangle$, and $\langle h_a, h_b \rangle$ is decided during construct by S . The l 0thellos may share the same $\langle h_a, h_b \rangle$ and same edges in G . However, the bitmaps in different 0thellos are different.

Parallelized execution of POG operations

We present how to execute the 0thello operations in parallel on l 0thellos in the same POG. To derive the $\tau(k)$ value of an l -POG, naïvely querying all l 0thellos requires $2l$ memory accesses. Our approach only needs 2 memory accesses to query an l -POG.

An l -POG query structure includes l, m, h_a, h_b and two vectors A and B . Each of A and B contains m l -bit integers. POG uses A and B to store the bitmaps in the bitmaps a_1, a_2, \dots, a_l and b_1, b_2, \dots, b_l . Let $A[i]$ be the i -th element of A . The t -th least significant bit of $A[i]$ is $a_t[i]$, where a_t is a bitmap of the t -th 0thello query structure. B is defined similarly. The t -th least significant bit of the l -bit integer value $A[h_a(k)] \oplus B[h_b(k)]$ is $a_t[h_a(k)] \oplus b_t[h_b(k)] = \tau_t(k)$. Hence, $\tau(k)$ can be computed by:

$$\tau(k) = A[h_a(k)] \oplus B[h_b(k)]$$

When l is not larger than the word size of the platform, each l -POG query only requires two memory accesses for fetching $A[i]$ and $B[j]$. The arithmetic operation includes computing the hash functions and the XOR value. On 64-bit platforms, a 64-POG is sufficient to support name switching with 2^{64} actions. Even if $l = 8$, it supports 255 ports plus a packet drop action.

This optimization applies to all other operations as well. All 0thello operations can be decomposed into two steps: (1) modifications on G and (2) operations on some bits in a and b . In an l -POG, the l 0thellos share the same G and the first step is only executed once for all l 0thellos. In the second step, the indexes of changed bits in a and b only depend on G . These operations on all l bits with the same index can be executed in parallel. For example, an operation on $a_l[i], a_{l-1}[i], \dots, a_1[i]$ can be implemented using one arith-

metic operation on l -bit integer $A[i]$. Therefore, the expected time cost of each name addition, deletion, or set change operation is only $O(1)$ instead of $O(l)$. The time complexity of construction of POG is still $O(n)$.

In summary, all XOR actions on elements in a and b of l 0thellos of the same POG, can be executed by one l -bit XOR operation. Hence, all operations of an l -POG, including query, construct, alter, addX, addY, and delete, have the same time cost as those of a single 0thello. The memory cost of an l -POG query structure is $2lm + O(1)$. Since l is a small constant and $l \leq 8$ in most cases, the memory space is still $O(n)$.

Network-wide shared bipartite graph. For some networks that require every switch to store all destination names, such as Ethernet, the name set S is identical for all switches in the network. Hence, all switches in the network may share the same G and $\langle h_a, h_b \rangle$. Constructing and updating the FIBs in all switches only require computing G once. This property significantly improves the controller scalability.

5.2 Selection of Hash functions

The hash function pair is critical for system efficiency. Ideally, h_a and h_b should be chosen from a family of fully random and uniform hash functions. Similar to the implementation of Cuckoo Hashing [34], we apply a function $H(k, \text{seed})$ to generate the hashes in our implementation. Here, H is a particular hashing method and seed is a 32-bit integer. We let $h_a(k) = H(k, \text{seed}_a)$ and $h_b(k) = H(k, \text{seed}_b)$. Thus, $\langle h_a, h_b \rangle$ is uniquely determined by a pair of integers $(\text{seed}_a, \text{seed}_b)$.

The proper hashing method $H()$ is platform-dependent. In our experiments, Concise uses the CRC32c function for stronger and faster hash results. Our evaluation shows that CRC32c demonstrates desirable performance in practice. Using a few arithmetic instructions, the 32-bit CRC32c value can be effectively mapped to $\{0, 1, \dots, 2^x - 1\}$ for some integer x .

5.3 FIB Update and Concurrency Control

We assume that there is one logically centralized controller in the network. Upon network dynamics, the controller computes the POGs for a number of switches and update the query structures in the switches by FIB update messages using a standard SDN API. If m, h_a, h_b do not change during the update, an update message only contains a list of elements to be modified in A and B . Otherwise, it contains the full query structure of l -POG $\langle m, h_a, h_b, A, B \rangle$.

After receiving an FIB update message, a Concise switch modifies its POG query structure. Instead of locks, Concise uses simple bit vectors to prevent read-write conflicts in the query structure. Experimental results show that the concurrency control mechanism only has a negligible impact to the network performance.

Concurrency requirements. Let A, B be the two vectors of the query structure before an update and A', B' be the ones after the update. For a name k that exists in the FIB before

and after the update, suppose $i = h_a(k)$ and $j = h_b(k)$, both $A[i] \oplus B[j]$ and $A'[i] \oplus B'[j]$ are considered as correct actions, although they may be different. Note that, when $A[i] = A'[i]$, the values $A'[i] \oplus B[j]$ and $A[i] \oplus B'[j]$ are both correct query results. Inconsistency only happens when both $A[i]$ and $B[j]$ are changed during the update.

Concurrency control design. Concise observes whether the vector A is being modified. For a query for name k , if an update that affects $A[i]$ is being executed, Concise does not execute the query until the update finishes. Concise maintains two bit vectors D_1 and D_2 for concurrency control. All bits in D_1 and D_2 are set to 0 during the initialization. Each index i ($0 \leq i < m$) corresponds to an index $p(i)$ in D_1 and D_2 . The lengths of D_1 and D_2 are set to 512 bits and $p(i) = i \bmod 512$.

Before an update of the POG that will change some elements of A , Concise flips the corresponding bits in D_1 , i.e., change 0s to 1s and 1s to 0s. After the update, it flips the bits with same indexes in D_2 . For any index i , when Concise observes $D_1[p(i)] \neq D_2[p(i)]$, there must be no ongoing update that affects $A[i]$. Note that even if a bit index corresponds to multiple elements that are changed in an update, the bit is only flipped once.

Query procedure. The query procedure for name k includes the following three steps. (1) Fetch the bit $\delta_2 = D_2[p(i)]$. (2) Fetch the value of $A[i]$ and $B[j]$. (3) Fetch $\delta_1 = D_1[p(i)]$. If $\delta_2 = \delta_1$, compute $A[i] \oplus B[j]$ and return it as the query result. Otherwise, $\delta_2 \neq \delta_1$, we know that the POG is currently being updated and the update affects $A[i]$. The query for k will stop and be put to a latter place of the query event queue.

Here, the order of flipping $D_1[p(i)]$ and $D_2[p(i)]$ during an update and the order of getting their values during a query are different. Any updates that affect $A[i]$ and start during an query must result in $\delta_2 \neq \delta_1$.

The above procedures of update and query should be executed in the given explicit order. This can be specified by compiler reorder barriers on strong memory model platforms such as x86_64, or fence instructions on weak memory model platforms such as ARM.

6. IMPLEMENTATION AND EVALUATION

We implement Concise on three platforms and conduct extensive experiments to evaluate its performance.

6.1 Implementation Platforms

1. Memory-mode. We implement the POG query and control structures, running on different cores of a desktop computer. In addition, we use a discrete-event simulator to simulate other data plane functions such as queuing. The memory-mode experiments are used to compare the performance of the algorithms and data structures. They demonstrate the maximum lookup speed that Concise is able to achieve on a computing device by eliminating the I/O overhead.

2. Click Modular Router [25] is an architecture for build-

ing configurable routers. We implement an Concise prototype on Click. It is able to serve as a real switch that forwards data packets.

3. Intel Data Plane Development Kit (DPDK) [17] is widely used in fast data plane designs [15, 46]. We use a virtualized environment to squeeze both the traffic generator and the forwarding application on a same physical machine. This prototype is able to serve as a real switch that forwards data packets.

Evaluation Platform. We conduct all experiments on a commodity desktop computer equipped with one Core i7-4770 CPU (4 physical cores @ 3.4 GHz, 8 MB L3 Cache shared by 8 logical cores) and 16 GB memory (Dual channel DDR3 1600MHz). Unlike CuckooSwitch [47] and ScaleBricks [46] that are specifically designed and optimized for a high-end many-core workstation or cluster, Concise is designed as a portable solution and we perform no platform-specific optimization.

6.2 Methodology

We compare Concise with three approaches for name switching: (1) Cuckoo hashing [34] (used in CuckooSwitch [47] and ScaleBricks [46]), (2) BUFFALO [43], and (3) Orthogonal Bloom filters. CuckooSwitch [34] is optimized for a specific platform with 16 cores and 40 MB cache. ScaleBricks [46] is designed for a high performance server cluster. We are not able to repeat their experiments on commodity desktop computers. Instead, we compare Concise with Cuckoo hashing, the FIB design used in [47] and [46], by reusing the code from the public repository of CuckooSwitch. BUFFALO does not always return correct forwarding actions. The false positive rate is set to at most 0.01%. We also implement a new technique called Orthogonal Bloom filters (OBFs) for comparison. It uses a Bloom filter to replace an Othello for classification of two sets X and Y : all names in X hit the Bloom filter. The false positive rate is also set to at most 0.01%. The other design of OBFs is similar to Concise.

We do not include SetSep [13, 46] in this section although it shares some similarity to Othello. SetSep is not proposed as the FIB structure in [46]. Instead, it is used as the global partition table to distribute the entire FIB to different nodes in a cluster. There is no explicit update algorithm for SetSep in either [13] or [46]. Hence, it is not suitable to implement SetSep and compare it with other FIB designs. However, we still implement a static version of SetSep and present some results in Sec. 7.2.

6.2.1 Data plane performance metrics

The following metrics characterize the performance of the Concise query structure in switches.

Memory cost: the size of memory to store a FIB.

MAQ: the maximum number of memory accesses per query. During each memory access, a word (64-bit data for 64-bit systems) is transmitted from memory to the CPU. It is used

to characterize the time cost of a query.

Query throughput: the number of queries that a FIB is able to process per second.

Query throughput under update: the query throughput measured when the FIB is being updated. It reflects the effectiveness of the concurrency control mechanism.

Processing delay: the processing delay of the query structure for a packet. A data plane device maintains a queue of packets when packet arrival rate exceeds the query throughput of the FIB. The processing delay reflects the ability of the data plane to process burst traffic. We use an event-based simulator to simulate the processing delay of both POG and Cuckoo hashing under real traffic trace. The traffic trace is replayed in 100x speed to simulate large traffic.

6.2.2 Control plane performance metrics

The following metrics characterize the performance of the Concise control structure in the controller.

Construction time: the time to construct a FIB. Note that, for some networks in which G is shared among all switch FIBs such as Ethernet, not every FIB requires the entire construction time. Once G is determined, it can be reused for all switches.

Update throughput: the number of updates can be processed by the control structure per second. Here, an update may be adding a name, deleting a name, or changing the forwarding action of a name.

6.2.3 LFSR name generator.

In the experiments, a series of query packets with different names were generated and fetched by the FIB. One straightforward approach is to feed the FIB with publicly available traffic trace. However, the time for transmitting the data from the physical memory to the cache is too large compared to the FIB query time. Hence, to conduct more accurate measurement, we use a linear feedback shift register (LFSR) to generate the names. LFSR is able to generate long pseudo-random number sequences efficiently with a very small memory overhead. One LFSR generates about 200M names per second on our platform. In addition, we provide event-based simulation using real traffic data to study the processing delay on Concise.

In fact, LFSR gives no favor to Concise in these evaluations because the names are generated in a round-robin scenario. The names are evenly distributed, which provides the minimum cache hit ratio. LFSR traffic is actually the *worst* traffic for Concise. On the contrary, in denial-of-service attack traffic, the queries concentrate on one or few names, and they always hit the cache. Hence, the query throughput of Concise in DoS attack traffic may be higher than the value measured with LFSR traffic.

6.3 Memory-mode Evaluation

6.3.1 Data plane performance

Memory efficiency and MAQ. Table 1 shows the size of memory of different types of FIBs. We compute the memory cost used by POG, Cuckoo hash table, BUFFALO, and OBFs, for five types of names: MAC addresses, IPv4, IPv6, OpenFlow matching fields, and file names. They have different sizes as shown in the table. Here IP addresses are only used as examples of a name type. These FIBs are not designed for IP prefix matching. The number of actions for OpenFlow could be very large. We let the number of actions be 256 and 32,768 and compute the FIB size respectively.

For the $(2, 4)$ -Cuckoo hash table (referred as Cuckoo), we use the same settings as described in [47]. For BUFFALO, we assume the names are evenly distributed among the actions, which gives an advantage to it. The lengths of Bloom filters are computed using the algorithm and the recommended parameter $k_{max} = 8$ in [43].

The memory space used by Concise is significantly smaller than that of Cuckoo, BUFFALO, and OBFs. It is only determined by the number of names n and the number of actions, and is independent of the name lengths. Table 1 also shows the maximum number of memory accesses per query (MAQ) of these FIBs. A smaller MAQ indicates fewer data transferred from the memory to the CPU, which results in better query throughput. Concise always requires exactly two memory accesses per query. The other FIBs may have larger MAQ depending on the name length and number of actions.

Query throughput versus number of names. Fig. 3 shows the query throughput of Concise, Cuckoo, BUFFALO, and OBFs. The names are MAC addresses (48-bit). l is set to 8. Recall that $l = 8$ is sufficient for as many as 256 actions. We evaluate the query speed of all FIBs on the computer with 4 CPU cores and 8MB L3 Cache. When n is smaller than 2 million, the throughput of Concise is very high (> 400 M queries per second (Mqps)). It is because the memory required by Concise is smaller than the cache size. When $n \geq 2$ M, the throughput decreases but still around 100 Mqps. This indicates that if other resources (e.g., I/O and buffer) are not the bottleneck, Concise is able to forward more than 100M packets per second, which is more than 800Gbps for 1000-Byte packets. The query performance decreases as the size of the query structure exceeds the CPU cache size. We observe similar results when running the evaluation on other machines with different CPUs. On the other hand, Cuckoo has the highest throughput among the remaining three FIBs but its throughput is only about 20% to 50% of that of Concise. The results of Cuckoo are consistent with those presented by the original CuckooSwitch paper². Note that the measured time overhead includes that of query genera-

²The paper [47] showed a throughput 4.2x as high as our Cuckoo results on a high-end machine with two Xeon E5-2680 CPUs (16 cores and 40MB L3 cache). It is approximately 4x as powerful as the one used in our experiments.

FIB Example				Concise		Cuckoo		BUFFALO		OBFs	
Name	Type	# Names	# Actions	Mem	MAQ	Mem	MAQ	Mem	MAQ	Mem	MAQ
MAC (48 bits)		7×10^5	16	1M	2	5.62M	2	2.64M	8	7.36M	15
MAC (48 bits)		5×10^6	256	16M	2	40.15M	2	27.70M	32	112.06M	16
MAC (48 bits)		3×10^7	256	96M	2	321.23M	2	166.23M	32	672.34M	16
IPv4 (32 bits)		1×10^6	16	1.5M	2	4.27M	5	3.77M	8	10.52M	15
IPv6 (128 bits)		2×10^6	256	4M	2	34.13M	17	11.08M	32	44.82M	16
OpenFlow (356b)		3×10^5	256	1M	2	14.46M	48	1.67M	32	6.72M	16
OpenFlow (356b)		1.4×10^6	65536	8M	2	67.46M	48	18.21M	9216	66.60M	17
File name (varied)		359194	16	512K	2	19.32M	52	1.35M	8	5.47M	15

Table 1: Memory and query cost comparison of four FIBs. MAQ: maximum number of memory accesses per query.

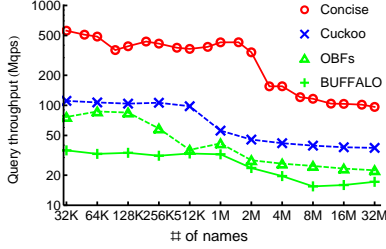


Figure 3: Query throughput versus number of names.

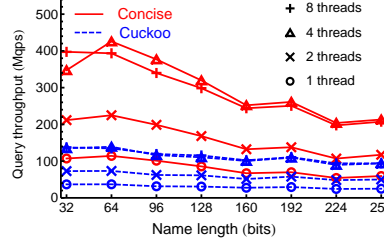


Figure 4: Query throughput versus name length

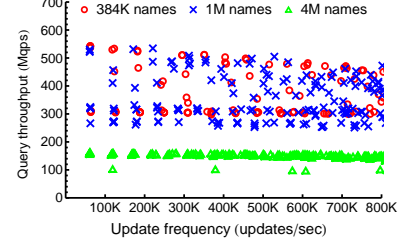


Figure 5: Concise query throughput under different update rates

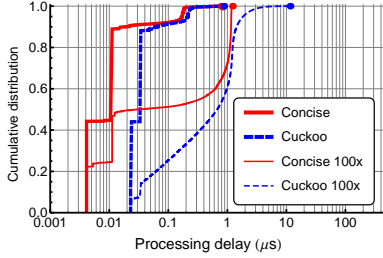


Figure 6: CDF of the processing delay of Concise and Cuckoo

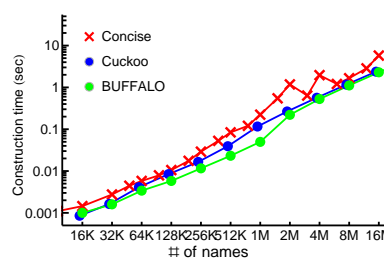


Figure 7: Construction time comparison among three FIBs

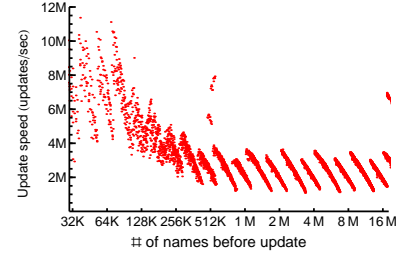


Figure 8: Concise update speed versus number of names

tion.³

Query throughput versus name length and number of CPU cores. Fig. 4 shows the query throughput using different lengths of names. Each FIB contains 256K names. As the length grows, the throughput of all types of Concise and Cuckoo FIBs decreases. Note that the memory size of Concise is independent of the name length. Hence, the throughput decrease of Concise is due to the increase of hashing time. One interesting observation is that, when the length is a multiple of 64 bits, the query throughput of Concise is slightly raised. This is mainly because the experiments are conducted on a 64-bit CPU. Meanwhile, as shown in Fig. 4, the query throughput grows approximately in proportional to the number of used threads, as long as the number of threads does not exceed the number of physical CPU cores of the platform.

³In the evaluation of 1M names, each query of Concise takes about 4.5 ns while generating a query takes 4.1 ns.

Query throughput during updates. Fig. 5 shows the throughput of Concise during updates, including name additions, deletions, and action changes. There is only very small decrease of query throughput even when the update frequency is as high as hundreds of thousands of names updated per second. For Concise with 4M names the throughput downgrade is negligible.

Processing delay. We conduct event-based simulations of packet processing on the data plane to study the process delay. We simulate a single-thread processor with two-level cache mechanism. The packets are processed in a first-come, first-served fashion. Each packet consists of the header and payload. The packets are put in a queue upon reception and wait to be processed by the processor. The processing delay of a packet is measured as the duration between the arrival to the FIB and finding the correct forwarding action. We measure the processing delay for real traffic data from the CAIDA Anonymized Internet Traces of December 2013 [1].

The average packet rate is about 210K packets per second. We plot the cumulative distribution of processing delays of Concise and Cuckoo in Fig. 6. Concise has smaller processing delay than Cuckoo before the 90th percentile, but they have similar tails. To study the processing delay under larger traffic volumes, we replay the CAIDA data 100x as fast as the original. The cumulative distributions of processing delay under 100x data rate are shown as the thin curves. The processing delay of Concise is clearly smaller than that of Cuckoo before the 60th percentile. After that, the two curves are similar, except that Cuckoo has a longer tail. Overall, the processing delay of Concise is very small ($< 1\mu s$) even under high data volumes.

6.3.2 Control plane performance

Construction time. Fig. 7 shows the average time to construct the query and control structures for one switch by varying the number of names n , where $l=8$. The construction time of Concise grows as the number of addresses increases. However, the construction time is very small. For 4M names, it takes only 1 second to construct the FIB. Note that the acyclic graph G can be reused for all other switches in the network. Hence, network-wide FIB construction only takes a few seconds.

We also compare the construction time of Concise to that of Cuckoo and BUFFALO. All these three FIBs can be constructed efficiently within similar construction time, though the time of Concise is larger than that of Cuckoo and BUFFALO.

Update speed. The update speed of Concise indicates its ability of maintaining a correct data plane under network dynamics. All types of network dynamics, including host and link changes, will be reflected as name additions, deletions, and action changes of the FIBs. Fig. 8 shows the update speed of Concise in the number of updates processed per second. We vary the number of names before update. In each experiment, the controller performs a number of updates and we measure the time to complete these updates. The number of updates per second is then computed. Each run of the experiment is shown as a point in the figure. When n becomes bigger the update speed decreases. In most cases, the update speed is very fast ($> 1M$ updates per second), which is sufficient for very large networks. In some rare cases, adding a new name may require reconstructing the POG. These cases happens with probability less than $\frac{3}{8n}E[|cc|]$, which is very small (about 1.3 parts per million when there are 1M names).

Communication overhead. To evaluate the overhead of Concise update messages sent from the controller to switches, we compute the entropy of the information included in update messages. The results are shown in Table 2. For $n = 3 \times 10^5$ names and 2^8 actions, the entropy is smaller than 80 bits. Hence, the message length is less than 10 Bytes for both name addition and action change. The update message length grows logarithmically with respect to either the number of names n or the number of actions. The communication

overhead of Concise is smaller than that of most OpenFlow operations.

	$n = 3 \times 10^5$ 2^8 actions	$n = 1.4 \times 10^6$ 2^{16} actions
Name addition	75.2	107.2
Action change	65.6	88.8

Table 2: Entropy of one update message in bits

6.4 Prototype Implementation and Evaluation

6.4.1 Implementation on Click

Click is a software architecture for building configurable routers. We implement a Concise prototype on Click. It receives packets from one inbound port and forwards each packet to one of its several outbound ports. Upon receiving a packet, it queries the POG using the address field of the packet, i.e., the name, and decides the outbound port of the packet. In addition, we implement the $(2,4)$ -Cuckoo hash table, OBFs, as well as the binary search mechanism on Click. Fig. 9 shows the forwarding throughput of all prototypes of different FIBs. The Click modules in each evaluation includes one traffic generator generating packets with valid 64-bit names, one switch that executes queries on the FIB, and packet counters connected to the egress ports of the switch. The experiments are conducted on one CPU core. The number of actions is 256.

Results show that Concise always has the highest throughput. When $n < 2M$, Concise is smaller than the cache size and the query throughput is about 2x as fast as Cuckoo and 4x as fast as OBFs. When $n \geq 2M$, the throughput of Concise is still the highest. Meanwhile, Concise uses much less memory, about 10% to 20% of that of Cuckoo, OBFs, and Binary.

6.4.2 Implementation with DPDK

We also build a prototype of Concise using Intel DPDK. Concise is implemented as a DPDK application on a general x86_64 Linux machine.

The prototype executes on the hardware Environment Abstraction Layer (EAL) provided by DPDK. It maintains a POG query structure. The query structure is initialized during boot up and can be updated upon network dynamics. The Concise prototype reads packets from the inbound ports, executes queries on the query structure, and then forwards each packet to the corresponding outbound port.

We implement both the traffic generator and FIB application on a same commodity computer using virtualization techniques. As shown in Fig. 10, we create a guest virtual machine (VM) on the host machine using KVM and Qemu to install Concise. The VM is equipped with four virtio-based virtual network interface cards. Linux TAP kernel virtual devices are attached to the virtio devices on the host side. The programs running on the host machine communicate with the guest VM via the Linux TAPs. On the

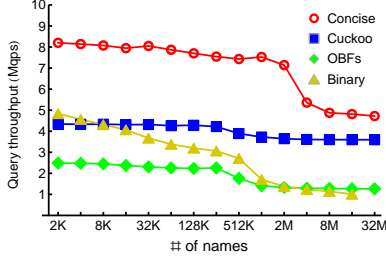


Figure 9: Forwarding throughput comparison on Click

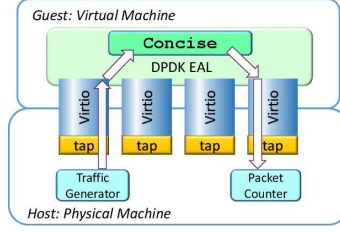


Figure 10: Concise prototype on DPDK

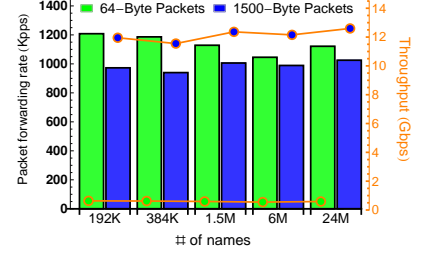


Figure 11: Performance of the Concise prototype on DPDK

host machine, we use a traffic generator program to send raw Ethernet packets to Concise running on the VM. The host machine receives the forwarded packets from Concise and counts the number of packets using default counters provided by the Linux system.

We measure the throughput of the Concise prototype with different numbers of names. The barchart in Fig. 11 shows that this prototype system is able to generate, forward, and receive more than 1M packets per second. Such performance is measured when Concise only uses one CPU core. The curve in Fig. 11 shows the forwarding throughput of the prototype system. The forward throughput is at least 12 Gbps for 1500-Byte Ethernet packets. Note that, in the previously results, the query throughput of Concise decreases as the FIB grows. The forwarding throughput does *not* decrease when the number of names grows. In addition, there is no significant difference between the packet rates of 64-Byte packets and 1500-Byte packets. This indicates the impact of Concise to the overall performance is so small that it is negligible compared to the other overheads. The bottleneck of this evaluation is on other parts of processing, e.g., the transmission of packets between the host machine and guest VM. We expect a much higher throughput on physical NICs.

7. DISCUSSION

7.1 Why Concise outperforms other designs?

Concise outperforms other FIB designs in multiple metrics. We summarize the merits as follows to explain the reasons of the performance gain.

1. Othello does *not* maintain a copy of the names in the query structure. The memory size of the query structure is much smaller than the other solutions. Concise demonstrates higher cache-hit rate, which leads to better performance on cache-based systems.
2. The query procedure does not contain any branches (e.g., if statements). This helps the CPU to predict and execute the instructions in the query procedure.

3. The Othello design allows us to build efficient and effective concurrency control mechanisms.

7.2 Concise versus Cuckoo and SetSep

Concise is essentially a classifier for n names, where each class represents a forwarding action. It does not store the names. Cuckoo uses a key-value store to represent the FIB, where the keys are the names and the values are the actions. It requires larger memory size than Concise because it needs to store all names and actions. Both Cuckoo and Concise compute two hash values per query. As shown in Table 1, when the name length is long, Cuckoo may need more than two memory accesses per query. Concise needs exactly two memory accesses despite of the name length and number of actions. Both Concise and Cuckoo rely on the properties of a graph where the edges represent the names: Concise requires the graph to be acyclic and Cuckoo requires the graph has no complex components [27]. However, they are completely different: the Concise design starts from an data structure that classifies names in two sets, while Cuckoo is a key-value store. Such difference results in the high query throughput and low memory usage of Concise. Meanwhile, the construction and update speed of Concise is slower than that of Cuckoo, but still fast enough for practical networks.

SetSep [13,46] shares some similar properties with Concise. For example, it does not store names. For an unknown name, SetSep also returns a meaningless result. SetSep is not used as the FIB design in [46]. Instead it is the separator to distribute FIBs to difference computers. We do not compare SetSep with other FIB designs in Sec. 6 because it does not include a design of the update algorithm in [13] and [46]. One limitation of SetSep is that the construction speed is slower than that of Concise and Cuckoo by more than an order of magnitude: 10 seconds for one single FIB of 1M names in our experiments. Unlike Cuckoo and Concise, one update result of SetSep cannot be reused in updates of other FIBs in the network wide. Hence, if SetSep is used as FIBs, the controller may have scalability problems. The memory cost of SetSep is about 50% to 70% of that of Concise according to the analysis in [46]. However, SetSep needs to

compute $1 + l$ hash values and read $2 + 2l$ values for each query. By applying the optimized memory structure, it still requires 3 memory accesses in many cases. Hence, its query speed is slower than that of Concise. We implement a static version of SetSep with 1.4M names and $l = 8$, using 2.19MB memory. Its query throughput is 211 Mqps using 4 threads. As comparison, Concise with same settings uses 4M memory and reaches 470 Mqps.

7.3 Eliminate Invalid Names

On Concise, querying a name that does not exist in the network may result in an arbitrary forwarding action. Compared to the forwarding table miss of Ethernet, which let the packets flood to all interfaces, Concise causes no flooding. However, we prefer to eliminate queries of invalid names. Operators may choose one or some of the following mechanisms for particular networks.

- At an ingress switch, every incoming packet should be checked by a filter to validate that its destination does exist in the network. One possible data structure to support this filter is a Cuckoo hash table [34]. This filter can be implemented as a network function running on the border of the network, and can be integrated with the firewall.
- Each query structure may maintain a Bloom filter to check whether the destination name is valid.
- In addition to the l -bit query results, also maintains the checksum for each name. Adding checksums will increase the memory size of Concise. For r -bit checksums, the overall memory cost of a query structure is $2(l + r)m + O(1)$. Note that as long as $l + r$ does not exceed the word length of the computing platform, the time overhead of all operations remain unchanged.

7.4 Example Use Case

Concise provides desired FIB properties for many current and future architecture designs that adopt flat names as mentioned in Sec. 1. We present a use case where it can be applied in a large enterprise network.

A large enterprise or data center network may include up to millions of end hosts and more VMs [20]. In these networks, internal flows contribute to the most bandwidth, which can be forwarded by Concise using destination names on layer 2. Hence, the destination of a packet in this network can only be either a host or a gateway. We require each host in the network voluntarily checks the validity of the packets before it sends out the packets. This can be easily achieved using software firewalls such as *iptables*, as all names of the hosts and gateways are known.

As of the gateway, we require it to execute two network functions: (1) For packets going out from the network, perform layer-three routing using the external IP of the destination. This is a basic function a router. (2) For packets going into the network, filter out all packets with invalid destinations. This can be implemented by a firewall. The packets

will be forwarded using the layer-two names of the destinations. In addition, we require all packets in the network to carry a time-to-live (TTL) value to prevent packets from being forwarded forever in case packets with invalid names pass the firewalls.

8. CONCLUSION

Concise is a portable FIB design for name switching. We propose a new data structure model to minimize memory cost of FIBs and move the construction and update functions to the SDN controller. Relying on minimal perfect hashing, we design a data structure *Othello* and use it to build the Concise system. We implement Concise using three platforms. According to our analysis and evaluation, Concise uses the smallest memory to achieve the fastest query speed among existing FIB solutions.

9. REFERENCES

- [1] The CAIDA UCSD Anonymized Internet Traces. http://www.caida.org/data/passive/passive_2013_dataset.xml.
- [2] D. G. Anderson, H. Balakrishnan, N. Feamster, T. Koponen, D. Moon, and S. Shenker. Accountable Internet Protocol (AIP). In *Proc. of ACM SIGCOMM*, 2008.
- [3] H. Asai and Y. Ohara. Poptrie: A Compressed Trie with Population Count for Fast and Scalable Software IP Routing Table Lookup. In *Proc. of ACM SIGCOMM*. ACM, aug 2015.
- [4] H. Balakrishnan, K. Lakshminarayanan, S. Ratnasamy, S. Shenker, I. Stoica, and M. Walfish. A layered naming architecture for the Internet. In *Proc. of ACM SIGCOMM*, 2004.
- [5] D. Belazzougui, P. Boldi, R. Pagh, and S. Vigna. Monotone minimal perfect hashing: searching a sorted table with $O(1)$ accesses. In *Proc. of ACM SODA*, 2009.
- [6] O. B. Bernard Chazelle, Joe Kilian, Ronitt Rubinfeld, Ayellet Tal. The Bloomier Filter: An Efficient Data Structure for Static Support Lookup Tables, 2004.
- [7] B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.
- [8] F. C. Botelho, R. Pagh, and N. Ziviani. Practical perfect hashing in nearly optimal space. *Information Systems*, 38(1):108–131, mar 2013.
- [9] F. C. Botelho, N. Wormald, and N. Ziviani. Cores of random r -partite hypergraphs. *Information Processing Letters*, 112(8-9):314–319, apr 2012.
- [10] M. Caesar, T. Condie, J. Kannan, K. Lakshminarayanan, I. Stoica, and S. Shenker. ROFL: Routing on Flat Labels. In *Proc. of ACM SIGCOMM*, 2006.
- [11] S. R. Chowdhury, M. F. Bari, R. Ahmed, and R. Boutaba. PayLess: A Low Cost Network Monitoring Framework for Software Defined Networks. In *Proc. of IEEE/IFIP NOMS*, 2014.
- [12] Z. J. Czech, G. Havas, and B. S. Majewski. An optimal algorithm for generating minimal perfect hash functions. *Information Processing Letters*, 43(5):257–264, 1992.
- [13] B. Fan, D. Zhou, H. Lim, M. Kaminsky, and D. G. Andersen. When cycles are cheap, some tables can be huge. In *Proc. of USENIX HotOS*. USENIX Association, may 2013.
- [14] B. A. Greenberg, J. R. Hamilton, S. Kandula, C. Kim, P. Lahiri, A. Maltz, P. Patel, S. Sengupta, A. Greenberg, N. Jain, and D. a. Maltz. VL2: a scalable and flexible data center network. *ACM SIGCOMM CCR*, 09:51–62, 2009.
- [15] S. Han, K. Jang, A. Panda, S. Palkar, D. Han, and S. Ratnasamy. SoftNIC: A Software NIC to Augment Hardware. Technical Report UCB/EECS-2015-155, EECS Department, University of California, Berkeley, may 2015.
- [16] J. Herzen, C. Westphal, and P. Thiran. Scalable routing easy as pie: A practical isometric embedding protocol. In *Proc. of IEEE ICNP*, 2011.
- [17] Intel. Data Plane Development Kit. <http://dpdk.org/>.

- [18] V. Jacobson, Smetters, D. K., J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard. Networking named content. In *Proc. of ACM CoNEXT*, 2009.
- [19] S. Jain, Y. Chen, S. Jain, and Z.-L. Zhang. VIRO: A Scalable, Robust and Name-space Independent Virtual Id ROuting for Future Networks. In *Proc. of IEEE INFOCOM*, 2011.
- [20] S. Jain and Others. B4: Experience with a Globally-Deployed Software Defined WAN. In *Proc. of ACM SIGCOMM*, 2013.
- [21] S. Janson and M. J. Luczak. Susceptibility in subcritical random graphs. *Journal of Mathematical Physics*, 49(12):125207, 2008.
- [22] Y. Kanizo, D. Hay, and I. Keslassy. Palette: Distributing tables in software-defined networks. In *Proc. of IEEE INFOCOM*, 2013.
- [23] P. Kazemian, G. Varghese, and N. McKeown. Header Space Analysis: Static Checking For Networks. In *Proc. of USENIX NSDI*, 2012.
- [24] C. Kim, M. Caesar, and J. Rexford. Floodless in seattle: a scalable ethernet architecture for large enterprises. In *Proc. of SIGCOMM*, 2008.
- [25] E. Kohler, R. Morris, and B. Chen. *The Click Modular Router*. PhD thesis, Massachusetts Institute of Technology, 2000.
- [26] T. Koponen, M. Chawla, B.-G. Chun, A. Ermolinskiy, K. H. Kim, S. Shenker, and I. Stoica. A data-oriented (and beyond) network architecture. *SIGCOMM CCR*, 37(4):181, 2007.
- [27] R. Kutzelnigg. Bipartite Random Graphs and Cuckoo Hashing. In *DMTCS Proc.*, number 1, 2006.
- [28] B. S. Majewski, N. Wormald, G. Havas, and Z. Czech. A Family of Perfect Hashing Methods. *The Computer Journal*, 39(6):547–554, jun 1996.
- [29] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner. OpenFlow: enabling innovation in campus networks. *SIGCOMM CCR*, 38(2):69–74, mar 2008.
- [30] M. Moradi, F. Qian, Q. Xu, Z. M. Mao, D. Bethea, and M. K. Reiter. Caesar: High-Speed and Memory-Efficient Forwarding Engine for Future Internet Architecture. In *Proc. of ACM/IEEE ANCS*, 2015.
- [31] R. Moskowitz, P. Nikander, P. Jokela, and T. Henderson. Host Identity Protocol. Technical report, 2008.
- [32] I. Müller, P. Sanders, R. Schulze, and W. Zhou. Retrieval and Perfect Hashing Using Fingerprinting. In *Experimental Algorithms*, pages 138–149. Springer, 2014.
- [33] D. Naylor and Others. XIA: Architecting a More Trustworthy and Evolvable Internet. *SIGCOMM CCR*, 44(3):50–57, 2014.
- [34] R. Pagh and F. F. Rodler. Cuckoo hashing. *Journal of Algorithms*, 51(2):122–144, may 2004.
- [35] C. Qian and S. Lam. ROME: Routing On Metropolitan-scale Ethernet. In *Proc. of IEEE ICNP*, 2012.
- [36] D. Raychaudhuri, K. Nagaraja, and A. Venkataramani. MobilityFirst: A Robust and Trustworthy MobilityCentric Architecture for the Future Internet. *Mobile Computer Communication Review*, 2012.
- [37] J. Saltzer. On the naming and binding of network destinations. RFC 1498, 1993.
- [38] D. Sampath, S. Agarwal, and J. J. Garcia-Luna-Aceves. Ethernet on AIR: Scalable Routing in Very Large Ethernet-based Networks. In *Proc. of IEEE ICDCS*, 2010.
- [39] A. Singla, P. B. Godfrey, K. Fall, G. Iannaccone, and S. Ratnasamy. Scalable Routing on Flat Names. In *Proc. of ACM CoNEXT*, 2010.
- [40] B. Stephens, A. Cox, W. Felter, C. Dixon, and J. Carter. PAST: Scalable Ethernet for Data Centers. In *Proc. of ACM CoNEXT*, 2012.
- [41] Y. Wang, Y. Zu, T. Zhang, K. Peng, Q. Dong, B. Liu, W. Meng, H. Dai, X. Tian, Z. Xu, H. Wu, and D. Yang. Wire speed name lookup: a GPU-based approach. *Proc. of USENIX NSDI*, 2013.
- [42] T. Yang, G. Xie, Y. Li, Q. Fu, A. X. Liu, Q. Li, and L. Mathy. Guarantee IP Lookup Performance with FIB Explosion. In *Proc. of ACM SIGCOMM*, 2014.
- [43] M. Yu, A. Fabrikant, and J. Rexford. BUFFALO: Bloom filter forwarding architecture for large organizations. In *Proc. of ACM CoNEXT*, 2009.
- [44] M. Yu, J. Rexford, M. J. Freedman, and J. Wang. Scalable flow-based networking with DIFANE. In *Proc. of ACM SIGCOMM*, 2010.
- [45] L. Zhang, D. Estrin, J. Burke, V. Jacobson, J. D. Thornton, D. K. Smetters, B. Zhang, G. Tsudik, D. Massey, C. Papadopoulos, T. Abdelzaher, L. Wang, P. Crowley, and E. Yeh. Named data networking (ndn) project. *NDN Tech. report*, 2010.
- [46] D. Zhou, B. Fan, H. Lim, D. G. Anderson, M. Kaminsky, M. Mitzenmacher, R. Wang, and A. Singh. Scaling Up Clustered Network Appliances with ScaleBricks. In *Proc. of ACM SIGCOMM*, 2015.
- [47] D. Zhou, B. Fan, H. Lim, M. Kaminsky, and D. G. Anderson. Scalable, High Performance Ethernet Forwarding with CuckooSwitch. In *Proc. of ACM CoNEXT*, 2013.